17 Random Variables and Distributions

Thus far, we have focused on probabilities of events. For example, we computed the probability that you win the Monty Hall game, or that you have a rare medical condition given that you tested positive. But, in many cases we would like to more more. For example, *how many* contestants must play the Monty Hall game until one of them finally wins? *How long* will this condition last? *How much* will I lose gambling with strange dice all night? To answer such questions, we need to work with random variables.

17.1 Definitions and Examples

Definition 17.1.1. A *random variable* R on a probability space is a total function whose domain is the sample space.

The codomain of R can be anything, but will usually be a subset of the real numbers. Notice that the name "random variable" is a misnomer; random variables are actually functions!

For example, suppose we toss three independent[1], unbiased coins. Let C be the number of heads that appear. Let $M = 1$ if the three coins come up all heads or all tails, and let $M = 0$ otherwise. Every outcome of the three coin flips uniquely determines the values of C and M. For example, if we flip heads, tails, heads, then $C = 2$ and $M = 0$. If we flip tails, tails, tails, then $C = 0$ and $M = 1$. In effect, C counts the number of heads, and M indicates whether all the coins match.

Since each outcome uniquely determines C and M, we can regard them as functions mapping outcomes to numbers. For this experiment, the sample space is

$$\mathcal{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \leftarrow$$

and C is a function that maps each outcome in the sample space to a number as

[1] Going forward, when we talk about flipping independent coins, we will assume that they are *mutually* independent.

follows:

$$\begin{aligned} C(HHH) &= 3 & C(THH) &= 2 \\ C(HHT) &= 2 & C(THT) &= 1 \\ C(HTH) &= 2 & C(TTH) &= 1 \\ C(HTT) &= 1 & C(TTT) &= 0. \end{aligned}$$

Similarly, M is a function mapping each outcome another way:

$$\begin{aligned} M(HHH) &= 1 & M(THH) &= 0 \\ M(HHT) &= 0 & M(THT) &= 0 \\ M(HTH) &= 0 & M(TTH) &= 0 \\ M(HTT) &= 0 & M(TTT) &= 1. \end{aligned}$$

So C and M are random variables.

17.1.1 Indicator Random Variables

An *indicator random variable* is a random variable that maps every outcome to either 0 or 1. Indicator random variables are also called *Bernoulli variables*. The random variable M is an example. If all three coins match, then $M = 1$; otherwise, $M = 0$.

Indicator random variables are closely related to events. In particular, an indicator random variable partitions the sample space into those outcomes mapped to 1 and those outcomes mapped to 0. For example, the indicator M partitions the sample space into two blocks as follows:

$$\underbrace{HHH \quad TTT}_{M=1} \quad \underbrace{HHT \quad HTH \quad HTT \quad THH \quad THT \quad TTH}_{M=0}.$$

In the same way, an event E partitions the sample space into those outcomes in E and those not in E. So E is naturally associated with an indicator random variable, I_E, where $I_E(w) = 1$ for outcomes $w \in E$ and $I_E(w) = 0$ for outcomes $w \notin E$. Thus, $M = I_E$ where E is the event that all three coins match.

17.1.2 Random Variables and Events

There is a strong relationship between events and more general random variables as well. A random variable that takes on several values partitions the sample space into several blocks. For example, C partitions the sample space as follows:

$$\underbrace{TTT}_{C=0} \quad \underbrace{TTH \quad THT \quad HTT}_{C=1} \quad \underbrace{THH \quad HTH \quad HHT}_{C=2} \quad \underbrace{HHH}_{C=3}.$$

Each block is a subset of the sample space and is therefore an event. Thus, we can regard an equation or inequality involving a random variable as an event. For example, the event that $C = 2$ consists of the outcomes THH, HTH, and HHT. The event $C \leq 1$ consists of the outcomes TTT, TTH, THT, and HTT.

Naturally enough, we can talk about the probability of events defined by properties of random variables. For example,

$$\begin{aligned}\Pr[C = 2] &= \Pr[THH] + \Pr[HTH] + \Pr[HHT] \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{3}{8}.\end{aligned}$$

As another example:

$$\begin{aligned}\Pr[M = 1] &= \Pr[TTT] + \Pr[HHH] \\ &= \frac{1}{8} + \frac{1}{8} \\ &= \frac{1}{4}.\end{aligned}$$

17.1.3 Functions of Random Variables

Random variables can be combined to form other random variables. For example, suppose that you roll two unbiased, independent 6-sided dice. Let D_i be the random variable denoting the outcome of the ith die for $i = 1, 2$. For example,

$$\Pr[D_1 = 3] = 1/6.$$

Then let $T = D_1 + D_2$. T is also a random variable and it denotes the sum of the two dice. For example,

$$\Pr[T = 2] = 1/36$$

and

$$\Pr[T = 7] = 1/6.$$

Random variables can be combined in complicated ways, as we will see in Chapter 19. For example,

$$Y = e^T$$

is also a random variable. In this case,

$$\Pr[Y = e^2] = 1/36$$

and

$$\Pr[Y = e^7] = 1/6.$$

17.1.4 Conditional Probability

Mixing conditional probabilities and events involving random variables creates no new difficulties. For example, $\Pr[C \geq 2 \mid M = 0]$ is the probability that at least two coins are heads ($C \geq 2$) given that not all three coins are the same ($M = 0$). We can compute this probability using the definition of conditional probability:

$$\begin{aligned}
\Pr[C \geq 2 \mid M = 0] &= \frac{\Pr[C \geq 2 \cap M = 0]}{\Pr[M = 0]} \\
&= \frac{\Pr[\{THH, HTH, HHT\}]}{\Pr[\{THH, HTH, HHT, HTT, THT, TTH\}]} \\
&= \frac{3/8}{6/8} \\
&= \frac{1}{2}.
\end{aligned}$$

The expression $C \geq 2 \cap M = 0$ on the first line may look odd; what is the set operation $\cap$ doing between an inequality and an equality? But recall that, in this context, $C \geq 2$ and $M = 0$ are *events*, and so they are *sets* of outcomes.

17.1.5 Independence

The notion of independence carries over from events to random variables as well. Random variables R_1 and R_2 are *independent* iff for all x_1 in the codomain of R_1, and x_2 in the codomain of R_2 for which $\Pr[R_2 = X_2] > 0$, we have:

$$\Pr[R_1 = x_1 \mid R_2 = x_2] = \Pr[R_1 = x_1].$$

As with events, we can formulate independence for random variables in an equivalent and perhaps more useful way: random variables R_1 and R_2 are independent if for all x_1 and x_2

$$\Pr[R_1 = x_1 \cap R_2 = x_2] = \Pr[R_1 = x_1] \cdot \Pr[R_2 = x_2].$$

For example, are C and M independent? Intuitively, the answer should be "no". The number of heads, C, completely determines whether all three coins match; that is, whether $M = 1$. But, to verify this intuition, we must find some $x_1, x_2 \in \mathbb{R}$ such that:

$$\Pr[C = x_1 \cap M = x_2] \neq \Pr[C = x_1] \cdot \Pr[M = x_2].$$

One appropriate choice of values is $x_1 = 2$ and $x_2 = 1$. In this case, we have:

$$\Pr[C = 2 \cap M = 1] = 0$$

and

$$\Pr[M = 1] \cdot \Pr[C = 2] = \frac{1}{4} \cdot \frac{3}{8} \neq 0.$$

The first probability is zero because we never have exactly two heads ($C = 2$) when all three coins match ($M = 1$). The other two probabilities were computed earlier.

On the other hand, let F be the indicator variable for the event that the first flip is a Head, so

$$\text{“}F = 1\text{”} = \{HHH, HTH, HHT, HTT\}.$$

Then F is independent of M, since

$$\Pr[M = 1] = 1/4 = \Pr[M = 1 \mid F = 1] = \Pr[M = 1 \mid F = 0]$$

and

$$\Pr[M = 0] = 3/4 = \Pr[M = 0 \mid F = 1] = \Pr[M = 0 \mid F = 0]$$

This example is an instance of a simple lemma:

Lemma 17.1.2. *Two events are independent iff their indicator variables are independent.*

As with events, the notion of independence generalizes to more than two random variables.

Definition 17.1.3. Random variables $R_1, R_2, \ldots, R_n$ are *mutually independent* iff

$$\begin{aligned}&\Pr[R_1 = x_1 \cap R_2 = x_2 \cap \cdots \cap R_n = x_n] \\ &= \Pr[R_1 = x_1] \cdot \Pr[R_2 = x_2] \cdots \Pr[R_n = x_n].\end{aligned}$$

for all $x_1, x_2, \ldots, x_n$.

A consequence of Definition 17.1.3 is that the probability that any *subset* of the variables takes a particular set of values is equal to the product of the probabilities that the individual variables take their values. Thus, for example, if $R_1, R_2, \ldots, R_{100}$ are mutually independent random variables, then it follows that:

$$\begin{aligned}&\Pr[R_1 = 7 \cap R_7 = 9.1 \cap R_{23} = \pi] \\ &\quad = \Pr[R_1 = 7] \cdot \Pr[R_7 = 9.1] \cdot \Pr[R_{23} = \pi].\end{aligned}$$

The proof is based on summing over all possible values for all of the other random variables.

17.2 Distribution Functions

A random variable maps outcomes to values. Often, random variables that show up for different spaces of outcomes wind up behaving in much the same way because they have the same probability of having any given value. Hence, random variables on different probability spaces may wind up having the same *probability density function*.

Definition 17.2.1. Let R be a random variable with codomain V. The *probability density function (pdf)* of R is a function $\mathrm{PDF}_R : V \to [0,1]$ defined by:

$$\mathrm{PDF}_R(x) ::= \begin{cases} \Pr[R = x] & \text{if } x \in \mathrm{range}(R) \\ 0 & \text{if } x \notin \mathrm{range}(R). \end{cases}$$

A consequence of this definition is that

$$\sum_{x \in \mathrm{range}(R)} \mathrm{PDF}_R(x) = 1.$$

This is because R has a value for each outcome, so summing the probabilities over all outcomes is the same as summing over the probabilities of each value in the range of R.

As an example, suppose that you roll two unbiased, independent, 6-sided dice. Let T be the random variable that equals the sum of the two rolls. This random variable takes on values in the set $V = \{2, 3, \ldots, 12\}$. A plot of the probability density function for T is shown in Figure 17.1: The lump in the middle indicates that sums close to 7 are the most likely. The total area of all the rectangles is 1 since the dice must take on exactly one of the sums in $V = \{2, 3, \ldots, 12\}$.

A closely-related concept to a PDF is the *cumulative distribution function (cdf)* for a random variable whose codomain is the real numbers. This is a function $\mathrm{CDF}_R : \mathbb{R} \to [0,1]$ defined by:

$$\mathrm{CDF}_R(x) = \Pr[R \leq x].$$

As an example, the cumulative distribution function for the random variable T is shown in Figure 17.2: The height of the ith bar in the cumulative distribution function is equal to the *sum* of the heights of the leftmost i bars in the probability

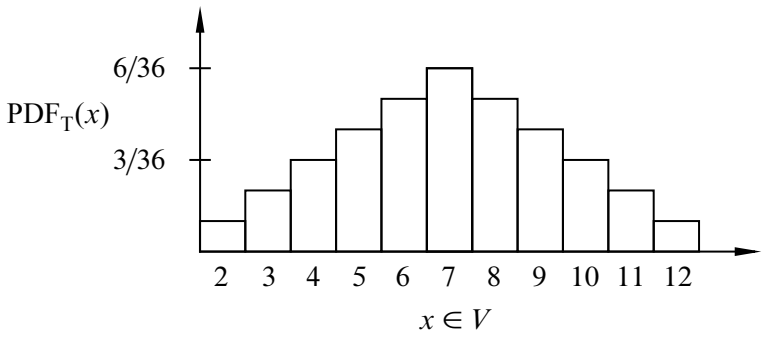

Figure 17.1 The probability density function for the sum of two 6-sided dice.

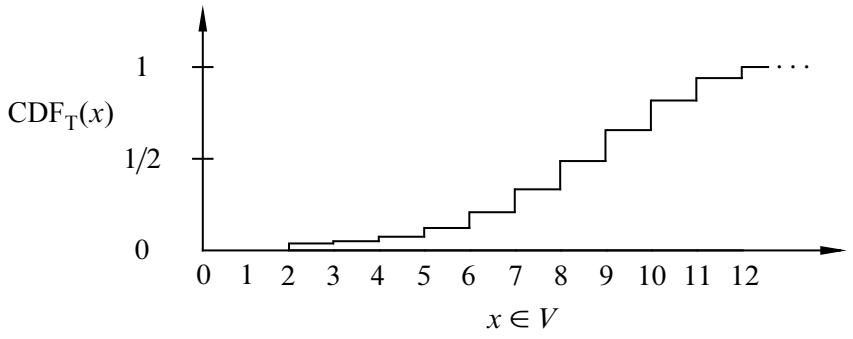

Figure 17.2 The cumulative distribution function for the sum of two 6-sided dice.

density function. This follows from the definitions of pdf and cdf:

$$\begin{aligned}\mathrm{CDF}_R(x) &= \Pr[R \le x] \\ &= \sum_{y \le x} \Pr[R = y] \\ &= \sum_{y \le x} \mathrm{PDF}_R(y).\end{aligned}$$

In summary, $\mathrm{PDF}_R(x)$ measures the probability that $R = x$ and $\mathrm{CDF}_R(x)$ measures the probability that $R \le x$. Both PDF_R and CDF_R capture the same information about the random variable R—you can derive one from the other—but sometimes one is more convenient.

One of the really interesting things about density functions and distribution functions is that many random variables turn out to have the *same* pdf and cdf. In other words, even though R and S are different random variables on different probability spaces, it is often the case that

$$\mathrm{PDF}_R = \mathrm{PDF}_S.$$

In fact, some pdfs are so common that they are given special names. For example, the three most important distributions in computer science are the *Bernoulli distribution*, the *uniform distribution*, and the *binomial distribution*. We look more closely at these common distributions in the next several sections.

17.3 Bernoulli Distributions

The Bernoulli distribution is the simplest and most common distribution function. That's because it is the distribution function for an indicator random variable. Specifically, the *Bernoulli distribution* has a probability density function of the form $f_p : \{0, 1\} \to [0, 1]$ where

$$\begin{aligned}f_p(0) &= p, \quad \text{and} \\ f_p(1) &= 1 - p,\end{aligned}$$

for some $p \in [0, 1]$. The corresponding cumulative distribution function is $F_p : \mathbb{R} \to [0, 1]$ where:

$$F_p(x) = \begin{cases} 0 & \text{if } x < 0 \\ p & \text{if } 0 \le x < 1 \\ 1 & \text{if } 1 \le x. \end{cases}$$

17.4 Uniform Distributions

17.4.1 Definition

A random variable that takes on each possible value with the same probability is said to be *uniform*. If the sample space is $\{1, 2, \ldots, n\}$, then the *uniform distribution* has a pdf of the form

$$f_n : \{1, 2, \ldots, n\} \to [0, 1]$$

where

$$f_n(k) = \frac{1}{n}$$

for some $n \in \mathbb{N}^+$. The cumulative distribution function is then $F_n : \mathbb{R} \to [0, 1]$ where

$$F_n(x) = \begin{cases} 0 & \text{if } x < 1 \\ k/n & \text{if } k \le x < k+1 \text{ for } 1 \le k < n \\ 1 & \text{if } n \le x. \end{cases}$$

Uniform distributions arise frequently in practice. For example, the number rolled on a fair die is uniform on the set $\{1, 2, \ldots, 6\}$. If $p = 1/2$, then an indicator random variable is uniform on the set $\{0, 1\}$.

17.4.2 The Numbers Game

Enough definitions—let's play a game! I have two envelopes. Each contains an integer in the range $0, 1, \ldots, 100$, and the numbers are distinct. To win the game, you must determine which envelope contains the larger number. To give you a fighting chance, we'll let you peek at the number in one envelope selected at random. Can you devise a strategy that gives you a better than 50% chance of winning?

For example, you could just pick an envelope at random and guess that it contains the larger number. But this strategy wins only 50% of the time. Your challenge is to do better.

So you might try to be more clever. Suppose you peek in one envelope and see the number 12. Since 12 is a small number, you might guess that that the number in the other envelope is larger. But perhaps we've been tricky and put small numbers in *both* envelopes. Then your guess might not be so good!

An important point here is that the numbers in the envelopes may *not* be random. We're picking the numbers and we're choosing them in a way that we think will defeat your guessing strategy. We'll only use randomization to choose the numbers if that serves our purpose, which is to make you lose!

Intuition Behind the Winning Strategy

Amazingly, there is a strategy that wins more than 50% of the time, regardless of what numbers we put in the envelopes!

Suppose that you somehow knew a number x that was in between the numbers in the envelopes. Now you peek in one envelope and see a number. If it is bigger than x, then you know you're peeking at the higher number. If it is smaller than x, then you're peeking at the lower number. In other words, if you know a number x between the numbers in the envelopes, then you are certain to win the game.

The only flaw with this brilliant strategy is that you do *not* know such an x. Oh well.

But what if you try to *guess* x? There is some probability that you guess correctly. In this case, you win 100% of the time. On the other hand, if you guess incorrectly, then you're no worse off than before; your chance of winning is still 50%. Combining these two cases, your overall chance of winning is better than 50%!

Informal arguments about probability, like this one, often sound plausible, but do not hold up under close scrutiny. In contrast, this argument sounds completely implausible—but is actually correct!

Analysis of the Winning Strategy

For generality, suppose that we can choose numbers from the set $\{0, 1, \ldots, n\}$. Call the lower number L and the higher number H.

Your goal is to guess a number x between L and H. To avoid confusing equality cases, you select x at random from among the half-integers:

$$\left\{\frac{1}{2},\ 1\frac{1}{2},\ 2\frac{1}{2},\ \ldots,\ n-\frac{1}{2}\right\}$$

But what probability distribution should you use?

The uniform distribution turns out to be your best bet. An informal justification is that if we figured out that you were unlikely to pick some number—say $50\frac{1}{2}$—then we'd always put 50 and 51 in the envelopes. Then you'd be unlikely to pick an x between L and H and would have less chance of winning.

After you've selected the number x, you peek into an envelope and see some number T. If $T > x$, then you guess that you're looking at the larger number. If $T < x$, then you guess that the other number is larger.

All that remains is to determine the probability that this strategy succeeds. We can do this with the usual four step method and a tree diagram.

Step 1: Find the sample space.
You either choose x too low ($< L$), too high ($> H$), or just right ($L < x < H$). Then you either peek at the lower number ($T = L$) or the higher number ($T = H$). This gives a total of six possible outcomes, as show in Figure 17.3.

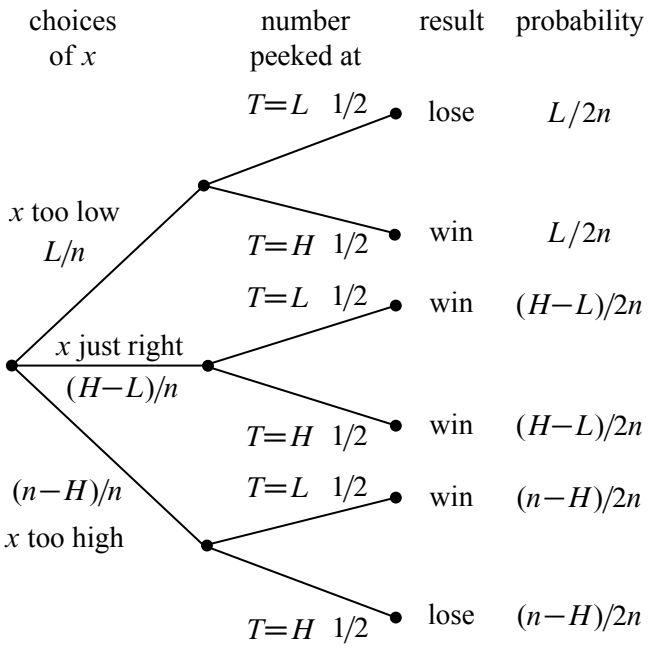

Figure 17.3 The tree diagram for the numbers game.

Step 2: Define events of interest.
The four outcomes in the event that you win are marked in the tree diagram.

Step 3: Assign outcome probabilities.
First, we assign edge probabilities. Your guess x is too low with probability L/n, too high with probability $(n - H)/n$, and just right with probability $(H - L)/n$. Next, you peek at either the lower or higher number with equal probability. Multiplying along root-to-leaf paths gives the outcome probabilities.

Step 4: Compute event probabilities.
The probability of the event that you win is the sum of the probabilities of the four outcomes in that event:

$$\begin{aligned}\Pr[\text{win}] &= \frac{L}{2n} + \frac{H-L}{2n} + \frac{H-L}{2n} + \frac{n-H}{2n} \\ &= \frac{1}{2} + \frac{H-L}{2n} \\ &\geq \frac{1}{2} + \frac{1}{2n}\end{aligned}$$

The final inequality relies on the fact that the higher number H is at least 1 greater than the lower number L since they are required to be distinct.

Sure enough, you win with this strategy more than half the time, regardless of the numbers in the envelopes! For example, if I choose numbers in the range $0, 1, \ldots, 100$, then you win with probability at least $\frac{1}{2} + \frac{1}{200} = 50.5\%$. Even better, if I'm allowed only numbers in the range $0, \ldots, 10$, then your probability of winning rises to 55%! By Las Vegas standards, those are great odds!

17.4.3 Randomized Algorithms

The best strategy to win the numbers game is an example of a *randomized algorithm*—it uses random numbers to influence decisions. Protocols and algorithms that make use of random numbers are very important in computer science. There are many problems for which the best known solutions are based on a random number generator.

For example, the most commonly-used protocol for deciding when to send a broadcast on a shared bus or Ethernet is a randomized algorithm known as *exponential backoff*. One of the most commonly-used sorting algorithms used in practice, called *quicksort*, uses random numbers. You'll see many more examples if you take an algorithms course. In each case, randomness is used to improve the probability that the algorithm runs quickly or otherwise performs well.

17.5 Binomial Distributions

17.5.1 Definitions

The third commonly-used distribution in computer science is the *binomial distribution*. The standard example of a random variable with a binomial distribution is the number of heads that come up in n independent flips of a coin. If the coin is

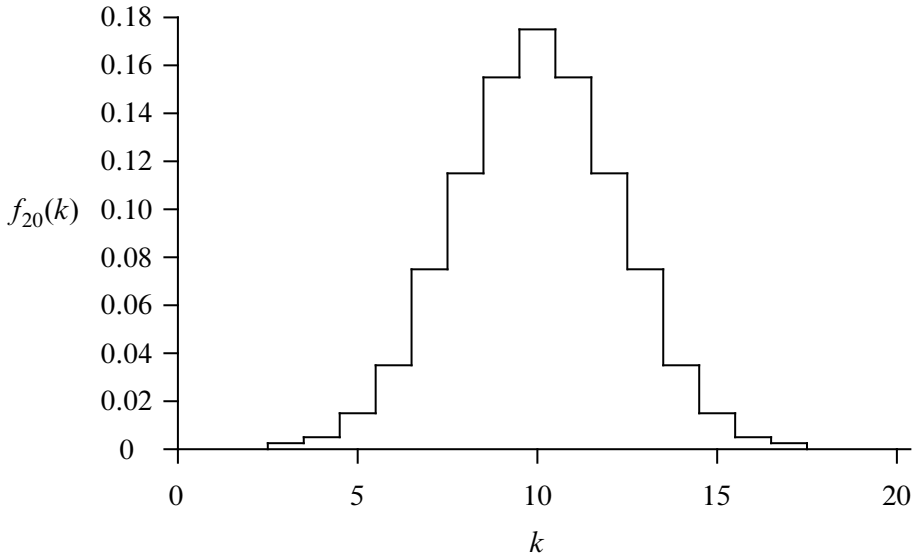

Figure 17.4 The pdf for the unbiased binomial distribution for $n = 20$, $f_{20}(k)$.

fair, then the number of heads has an *unbiased binomial distribution*, specified by the pdf

$$f_n : \{1, 2, \ldots, n\} \to [0, 1]$$

where

$$f_n(k) = \binom{n}{k} 2^{-n}$$

for some $n \in \mathbb{N}^+$. This is because there are $\binom{n}{k}$ sequences of n coin tosses with exactly k heads, and each such sequence has probability 2^{-n}.

A plot of $f_{20}(k)$ is shown in Figure 17.4. The most likely outcome is $k = 10$ heads, and the probability falls off rapidly for larger and smaller values of k. The falloff regions to the left and right of the main hump are called the *tails of the distribution*. We'll talk a lot more about these tails shortly.

The cumulative distribution function for the unbiased binomial distribution is $F_n : \mathbb{R} \to [0, 1]$ where

$$F_n(x) = \begin{cases} 0 & \text{if } x < 1 \\ \sum_{i=0}^{k} \binom{n}{i} 2^{-n} & \text{if } k \le x < k+1 \text{ for } 1 \le k < n \\ 1 & \text{if } n \le x. \end{cases}$$

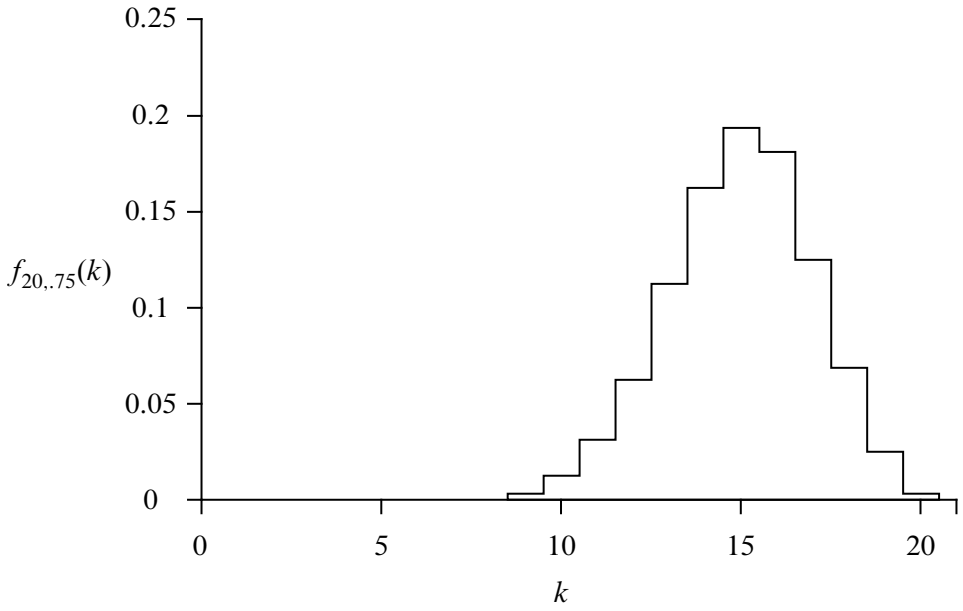

Figure 17.5 The pdf for the general binomial distribution $f_{n,p}(k)$ for $n = 20$ and $p = .75$.

The General Binomial Distribution

If the coins are biased so that each coin is heads with probability p, then the number of heads has a *general binomial density function* specified by the pdf

$$f_{n,p} : \{1, 2, \ldots, n\} \to [0, 1]$$

where

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

for some $n \in \mathbb{N}^+$ and $p \in [0, 1]$. This is because there are $\binom{n}{k}$ sequences with k heads and $n - k$ tails, but now the probability of each such sequence is $p^k(1-p)^{n-k}$.

For example, the plot in Figure 17.5 shows the probability density function $f_{n,p}(k)$ corresponding to flipping $n = 20$ independent coins that are heads with probability $p = 0.75$. The graph shows that we are most likely to get $k = 15$ heads, as you might expect. Once again, the probability falls off quickly for larger and smaller values of k.

The cumulative distribution function for the general binomial distribution is $F_{n,p} :\leftarrow \mathbb{R} \to [0,1]$ where

$$F_{n,p}(x) = \begin{cases} 0 & \text{if } x < 1 \\ \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} & \text{if } k \le x < k+1 \text{ for } 1 \le k < n \\ 1 & \text{if } n \le x. \end{cases} \tag{17.1}$$

17.5.2 Approximating the Probability Density Function

Computing the general binomial density function is daunting when k and n are large. Fortunately, there is an approximate closed-form formula for this function based on an approximation for the binomial coefficient. In the formula below, k is replaced by αn where α is a number between 0 and 1.

Lemma 17.5.1.

$$\binom{n}{\alpha n} \sim \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \tag{17.2}$$

and

$$\binom{n}{\alpha n} < \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \tag{17.3}$$

where $H(\alpha)$ *is the* entropy function[2]

$$H(\alpha) ::= \alpha \log\left(\frac{1}{\alpha}\right) + (1-\alpha)\log\left(\frac{1}{1-\alpha}\right)$$

Moreover, if $\alpha n > 10$ *and* $(1-\alpha)n > 10$*, then the left and right sides of Equation 17.2 differ by at most* 2%*. If* $\alpha n > 100$ *and* $(1-\alpha)n > 100$*, then the difference is at most* 0.2%*.*

The graph of H is shown in Figure 17.6.

Lemma (17.5.1) provides an excellent approximation for binomial coefficients. We'll skip its derivation, which consists of plugging in Theorem 9.6.1 for the factorials in the binomial coefficient and then simplifying.

Now let's plug Equation 17.2 into the general binomial density function. The probability of flipping αn heads in n tosses of a coin that comes up heads with

[2] $\log(x)$ means $\log_2(x)$.

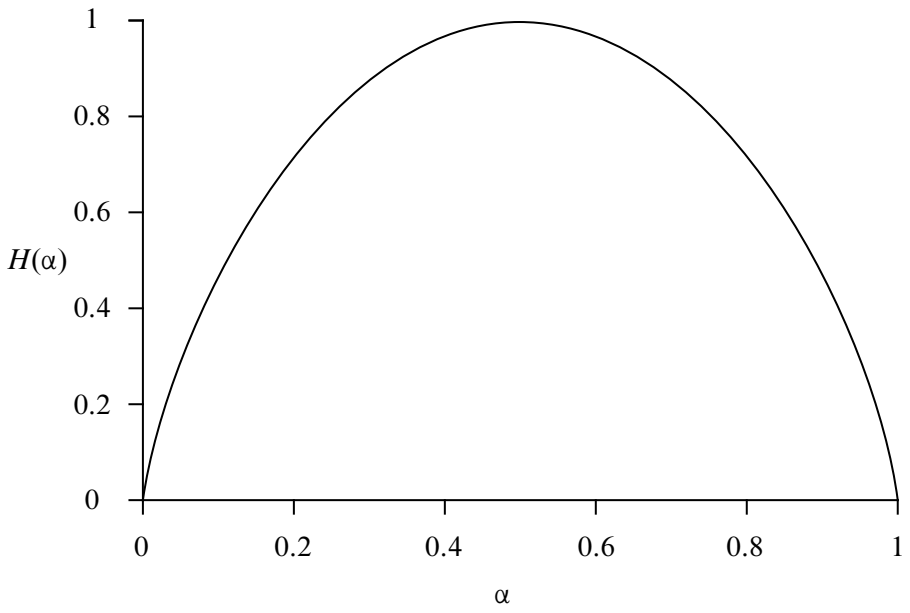

Figure 17.6 The Entropy Function

probability p is:

$$f_{n,p}(\alpha n) \sim \frac{2^{nH(\alpha)} p^{\alpha n} (1-p)^{(1-\alpha)n}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

$$= \frac{2^{n\left(\alpha \log\left(\frac{p}{\alpha}\right) + (1-\alpha)\log\left(\frac{1-p}{1-\alpha}\right)\right)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \qquad (17.4)$$

where the margin of error in the approximation is the same as in Lemma 17.5.1. From Equation 17.3, we also find that

$$f_{n,p}(\alpha n) < \frac{2^{n\left(\alpha \log\left(\frac{p}{\alpha}\right) + (1-\alpha)\log\left(\frac{1-p}{1-\alpha}\right)\right)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \qquad (17.5)$$

The formula in Equations 17.4 and 17.5 is as ugly as a bowling shoe, but it's useful because it's easy to evaluate. For example, suppose we flip a fair coin n times. What is the probability of getting *exactly* pn heads? Plugging $\alpha = p$ into Equation 17.4 gives:

$$f_{n,p}(pn) \sim \frac{1}{\sqrt{2\pi p(1-p)n}}.$$

Thus, for example, if we flip a fair coin (where $p = 1/2$) $n = 100$ times, the probability of getting exactly 50 heads is within 2% of 0.079, which is about 8%.

17.5.3 Approximating the Cumulative Distribution Function

In many fields, including computer science, probability analyses come down to getting small bounds on the tails of the binomial distribution. In a typical application, you want to bound the tails in order to show that there is very small probability that too many *bad* things happen. For example, we might like to know that it is very unlikely that too many bits are corrupted in a message, or that too many servers or communication links become overloaded, or that a randomized algorithm runs for too long.

So it is usually good news that the binomial distribution has small tails. To get a feel for their size, consider the probability of flipping at most 25 heads in 100 independent tosses of a fair coin.

The probability of getting at most αn heads is given by the binomial cumulative distribution function

$$F_{n,p}(\alpha n) = \sum_{i=0}^{\alpha n} \binom{n}{i} p^i (1-p)^{n-i}. \tag{17.6}$$

We can bound this sum by bounding the ratio of successive terms.

In particular, for $i \leq \alpha n$,

$$\begin{aligned}
\frac{\binom{n}{i-1} p^{i-1}(1-p)^{n-(i-1)}}{\binom{n}{i} p^i (1-p)^{n-i}} &= \frac{\frac{n!\,p^{i-1}(1-p)^{n-i+1}}{(i-1)!(n-i+1)!}}{\frac{n!\,p^i(1-p)^{n-i}}{i!(n-i)!}} \\
&= \frac{i(1-p)}{(n-i+1)p} \\
&\leq \frac{\alpha n(1-p)}{(n-\alpha n+1)p} \\
&\leq \frac{\alpha(1-p)}{(1-\alpha)p}.
\end{aligned}$$

This means that for $\alpha < p$,

$$\begin{aligned} F_{n,p}(\alpha n) &< f_{n,p}(\alpha n) \sum_{i=0}^{\infty} \left[\frac{\alpha(1-p)}{(1-\alpha)p}\right]^i \\ &= \frac{f_{n,p}(\alpha n)}{1 - \frac{\alpha(1-p)}{(1-\alpha)p}} \\ &= \left(\frac{1-\alpha}{1-\alpha/p}\right) f_{n,p}(\alpha n). \end{aligned} \tag{17.7}$$

In other words, the probability of at most αn heads is at most

$$\frac{1-\alpha}{1-\alpha/p}$$

times the probability of exactly αn heads. For our scenario, where $p = 1/2$ and $\alpha = 1/4$,

$$\frac{1-\alpha}{1-\alpha/p} = \frac{3/4}{1/2} = \frac{3}{2}.$$

Plugging $n = 100$, $\alpha = 1/4$, and $p = 1/2$ into Equation 17.5, we find that the probability of at most 25 heads in 100 coin flips is

$$F_{100,1/2}(25) < \frac{3}{2} \cdot \frac{2^{100\left(\frac{1}{4}\log(2) + \frac{3}{4}\log\left(\frac{2}{3}\right)\right)}}{\sqrt{75\pi/2}} \leq 3 \cdot 10^{-7}.$$

This says that flipping 25 or fewer heads is extremely unlikely, which is consistent with our earlier claim that the tails of the binomial distribution are very small. In fact, notice that the probability of flipping *25 or fewer* heads is only 50% more than the probability of flipping *exactly 25* heads. Thus, flipping exactly 25 heads is twice as likely as flipping any number between 0 and 24!

Caveat. The upper bound on $F_{n,p}(\alpha n)$ in Equation 17.7 holds only if $\alpha < p$. If this is not the case in your problem, then try thinking in complementary terms; that is, look at the number of tails flipped instead of the number of heads. In fact, this is precisely what we will do in the next example.

17.5.4 Noisy Channels

Suppose you are sending packets of data across a communication channel and that each packet is lost with probability $p = .01$. Also suppose that packet losses are independent. You need to figure out how much redundancy (or error correction) to

build into your communication protocol. Since redundancy is expensive overheard, you would like to use as little as possible. On the other hand, you never want to be caught short. Would it be safe for you to assume that in any batch of 10,000 packets, only 200 (or 2%) are lost? Let's find out.

The noisy channel is analogous to flipping $n = 10{,}000$ independent coins, each with probability $p = .01$ of coming up heads, and asking for the probability that there are at least αn heads where $\alpha = .02$. Since $\alpha > p$, we cannot use Equation 17.7. So we need to recast the problem by looking at the numbers of tails. In this case, the probability of tails is $p = .99$ and we are asking for the probability of at most αn tails where $\alpha = .98$.

Now we can use Equations 17.5 and 17.7 to find that the probability of losing 2% or more of the 10,000 packets is at most

$$\left(\frac{1-.98}{1-.98/.99}\right)\left(\frac{2^{10000\left(.98\log\left(\frac{.99}{.98}\right)+.02\log\left(\frac{.01}{.02}\right)\right)}}{\sqrt{2\pi(.98)(1-.98)10000}}\right) < 2^{-60}.$$

This is good news. It says that planning on at most 2% packet loss in a batch of 10,000 packets should be very safe, at least for the next few millennia.

17.5.5 Estimation by Sampling

Sampling is a very common technique for estimating the fraction of elements in a set that have a certain property. For example, suppose that you would like to know how many Americans plan to vote for the Republican candidate in the next presidential election. It is infeasible to ask every American how they intend to vote, so pollsters will typically contact n Americans selected at random and then compute the fraction of *those* Americans that will vote Republican. This value is then used as the *estimate* of the number of all Americans that will vote Republican. For example, if 45% of the n contacted voters report that they will vote Republican, the pollster reports that 45% of all Americans will vote Republican. In addition, the pollster will usually also provide some sort of qualifying statement such as

> "There is a 95% probability that the poll is accurate to within ± 4 percentage points."

The qualifying statement is often the source of confusion and misinterpretation. For example, many people interpret the qualifying statement to mean that there is a 95% chance that between 41% and 49% of Americans intend to vote Republican. But this is wrong! The fraction of Americans that intend to vote Republican is a fixed (and unknown) value p that is *not* a random variable. Since p is not a random variable, we cannot say anything about the probability that $.41 \leq p \leq .49$.

To obtain a correct interpretation of the qualifying statement and the results of the poll, it is helpful to introduce some notation.

Define R_i to be the indicator random variable for the ith contacted American in the sample. In particular, set $R_i = 1$ if the ith contacted American intends to vote Republican and $R_i = 0$ otherwise. For the purposes of the analysis, we will assume that the ith contacted American is selected uniformly at random (with replacement) from the set of all Americans.[3] We will also assume that every contacted person responds honestly about whether or not they intend to vote Republican and that there are only two options—each American intends to vote Republican or they don't. Thus,

$$\Pr[R_i = 1] = p \tag{17.8}$$

where p is the (unknown) fraction of Americans that intend to vote Republican.

We next define

$$T = R_1 + R_2 + \cdots + R_n$$

to be the number of contacted Americans who intend to vote Republican. Then T/n is a random variable that is the estimate of the fraction of Americans that intend to vote Republican.

We are now ready to provide the correct interpretation of the qualifying statement. The poll results mean that

$$\Pr\left[\,|T/n - p| \leq .04\right] \geq .95. \tag{17.9}$$

In other words, there is a 95% chance that the sample group will produce an estimate that is within ± 4 percentage points of the correct value for the overall population. So either we were "unlucky" in selecting the people to poll or the results of the poll will be correct to within ± 4 points.

How Many People Do We Need to Contact?

There remains an important question: how many people n do we need to contact to make sure that Equation 17.9 is true? In general, we would like n to be as small as possible in order to minimize the cost of the poll.

Surprisingly, the answer depends only on the desired *accuracy* and *confidence* of the poll and not on the number of items in the set being sampled. In this case, the desired accuracy is .04, the desired confidence is .95, and the set being sampled is the set of Americans. It's a good thing that n won't depend on the size of the set being sampled—there are over 300 million Americans!

[3] This means that someone could be contacted multiple times.

The task of finding an n that satisfies Equation 17.9 is made tractable by observing that T has a general binomial distribution with parameters n and p and then applying Equations 17.5 and 17.7. Let's see how this works.

Since we will be using bounds on the tails of the binomial distribution, we first do the standard conversion

$$\Pr\left[\left|T/n - p\right| \leq .04\right] = 1 - \Pr\left[\left|T/n - p\right| > .04\right]$$

We then proceed to upper bound

$$\begin{aligned}\Pr\left[\left|T/n - p\right| > .04\right] &\leq \Pr[T < (p - .04)n] + \Pr[T > (p + .04)n] \\ &= F_{n,p}((p - 0.4)n) + F_{n,1-p}((1 - p - .04)n). \quad (17.10)\end{aligned}$$

We don't know the true value of p, but it turns out that the expression on the righthand side of Equation 17.10 is maximized when $p = 1/2$ and so

$$\begin{aligned}\Pr\left[\left|T/n - p\right| > .04\right] &\leq 2F_{n,1/2}(.46n) \\ &< 2\left(\frac{1 - .46}{1 - (.46/.5)}\right) f_{n,1/2}(.46n) \\ &< 13.5 \cdot \frac{2^{n\left(.46\log\left(\frac{.5}{.46}\right) + .54\log\left(\frac{.5}{.54}\right)\right)}}{\sqrt{2\pi \cdot 0.46 \cdot 0.54 \cdot n}} \\ &< \frac{10.81 \cdot 2^{-.00462n}}{\sqrt{n}}. \quad (17.11)\end{aligned}$$

The second line comes from Equation 17.7 using $\alpha = .46$. The third line comes from Equation 17.5.

Equation 17.11 provides bounds on the confidence of the poll for different values of n. For example, if $n = 665$, the bound in Equation 17.11 evaluates to .04978 .... Hence, if the pollster contacts 665 Americans, the poll will be accurate to within ± 4 percentage points with at least 95% probability.

Since the bound in Equation 17.11 is exponential in n, the confidence increases greatly as n increases. For example, if $n = 6{,}650$ Americans are contacted, the poll will be accurate to within ± 4 points with probability at least $1 - 10^{-10}$. Of course, most pollsters are not willing to pay the added cost of polling 10 times as many people when they already have a confidence level of 95% from polling 665 people.

6.042J / 18.062J Mathematics for Computer Science
Fall 2010